Digital Gene Expression Signatures for Maize Development[1][W][OA]

Andrea L. Eveland, Namiko Satoh-Nagasawa[2], Alexander Goldshmidt, Sandra Meyer, Mary Beatty, Hajime Sakai, Doreen Ware[3], and David Jackson[3]*

Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724 (A.L.E., A.G., D.W., D.J.); DuPont Agricultural Biotechnology Experimental Station E353, Wilmington, Delaware 19880 (N.S.-N., H.S.); Pioneer Hi-Bred International, Johnston, Iowa 50131–1004 (S.M., M.B.); and United States Department of Agriculture-Agricultural Research Service, Robert W. Holley Center for Agriculture and Health, Ithaca, New York 14853 (D.W.)

Genome-wide expression signatures detect specific perturbations in developmental programs and contribute to functional resolution of key regulatory networks. In maize (*Zea mays*) inflorescences, mutations in the *RAMOSA* (*RA*) genes affect the determinacy of axillary meristems and thus alter branching patterns, an important agronomic trait. In this work, we developed and tested a framework for analysis of tag-based, digital gene expression profiles using Illumina's high-throughput sequencing technology and the newly assembled B73 maize reference genome. We also used a mutation in the *RA3* gene to identify putative expression signatures specific to stem cell fate in axillary meristem determinacy. The *RA3* gene encodes a trehalose-6-phosphate phosphatase and may act at the interface between developmental and metabolic processes. Deep sequencing of digital gene expression libraries, representing three biological replicate ear samples from wild-type and *ra3* plants, generated 27 million 20- to 21-nucleotide reads with frequencies spanning 4 orders of magnitude. Unique sequence tags were anchored to 3′-ends of individual transcripts by *Dpn*II and *Nla*III digests, which were multiplexed during sequencing. We mapped 86% of nonredundant signature tags to the maize genome, which associated with 37,117 gene models and unannotated regions of expression. In total, 66% of genes were detected by at least nine reads in immature maize ears. We used comparative genomics to leverage existing information from Arabidopsis (*Arabidopsis thaliana*) and rice (*Oryza sativa*) in functional analyses of differentially expressed maize genes. Results from this study provide a basis for the analysis of short-read expression data in maize and resolved specific expression signatures that will help define mechanisms of action for the *RA3* gene.

Genome-wide expression analyses provide essential building blocks for elucidating molecular function. Recent studies have highlighted the significance of high-throughput expression data, particularly with the integration of large, diverse data sets, in constructing biochemical and regulatory networks in silico (Levesque et al., 2006; Gutiérrez et al., 2007; Capaldi et al., 2008; Ramsey et al., 2008; Amit et al., 2009). Resolution of these networks is enhanced by increased sensitivity and specificity for transcript detection and by the availability of resources for a given species. For the model organism Arabidopsis (*Arabidopsis thaliana*), large-scale, community-generated expression data sets have been assembled and integrated into Web-based repositories (for review, see Bevan and Walsh, 2005; Brady and Provart, 2009). Interrogation of these data sets using systems approaches has identified key transcriptional regulators in various aspects of plant biology (Gutiérrez et al., 2008; Kaufmann et al., 2009; Pruneda-Paz et al., 2009). With the release of the first assembled maize (*Zea mays*) B73 reference genome sequence (Schnable et al., 2009) comes a need for comparable resources in maize. Comprehensive expression profiles for maize, as well as leveraging of existing information from comparative studies with other model plant species, are pivotal to fueling exploratory research of agriculturally important traits.

Over the past decade, since the first expression studies using microarrays, a major focus of the scientific community has been the accumulation and cataloging of genome-wide transcript data (Zimmermann et al., 2004; Toufighi et al., 2005; Goda et al., 2008; Barrett et al., 2009; Parkinson et al., 2009). Consistent with this trend, advances in array technology have progressively enhanced specificity, coverage, and the ability to address unique research questions. Genome-

[1] This work was supported by the U.S. Department of Agriculture (grant nos. NRICGP–2005–02402 to D.J. and 1907–21000–030–00 to D.W.) and by the National Science Foundation (grant no. DBI–0604923 to D.J. and Postdoctoral Fellowship in Biological Informatics award no. 0805655 to A.L.E.).

[2] Present address: Department of Biological Production, Laboratory of Plant Genetics and Breeding, Kaidobata-nishi 241-438, Shimo-Shinjyo Nakano, Akita City 010–0195, Japan.

[3] These authors contributed equally to the article.

* Corresponding author; e-mail jacksond@cshl.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: David Jackson (jacksond@cshl.edu).

[W] The online version of this article contains Web-only data.

[OA] Open Access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.110.159673

wide tiling arrays provide resolution up to the single nucleotide level and have been utilized to identify transcript variants (such as alternatively spliced transcripts), single-feature polymorphisms, and epigenetic marks (for review, see Gregory et al., 2008). Although microarray technologies continue to evolve, the recent emergence of deep-sequencing platforms has motivated the current digital age of functional genomics (Blow, 2009; Lister et al., 2009, Metzker, 2010). Next-generation technologies, such as those developed by Illumina (previously Solexa), 454 Life Sciences (Roche), and Applied Biosystems (ABI), can generate tens of thousands (Roche-454) to tens of millions (Illumina and ABI) of sequence reads, in parallel, with exceptional reproducibility (Li et al., 2008; Marioni et al., 2008; Simon et al., 2009). Adapting these technologies to genome-wide expression studies circumvents the inherent limitations of hybridization-based methods. For example, sequence-based methods require no prior knowledge of sequence and/or transcript composition and thus provide a potentially unbiased view of the transcriptome, which is not limited to fully sequenced genomes (Blencowe et al., 2009; Simon et al., 2009). In addition, sequencing technologies enable the resolution of transcript variants and novel mRNAs (Sultan et al., 2008; Marioni et al., 2008), minimize biases due to cross-hybridization (Tang et al., 2009), and provide quantitative measures of transcript abundance based on read count over a wide dynamic range ('t Hoen et al., 2008; Mortazavi et al., 2008; Sultan et al., 2008; Morrissy et al., 2009; Babbitt et al., 2010).

A number of studies have used next-generation sequencing technologies for genome-scale expression analyses in higher eukaryotes. Such approaches include whole-transcript sequencing and assembly (RNA-seq) using the long-read, 454 platform (Emrich et al., 2007a; Weber et al., 2007) and the massively parallel Illumina (Mortazavi et al., 2008; Wang et al., 2008, 2009) and ABI SOLiD (Tang et al., 2009) systems. While Illumina and ABI achieve a much greater depth of sequencing, read lengths are significantly shorter (typically 36–75 bases) compared with 454 (up to 500 bases). Alternatively, tag-based approaches target 3′-ends of transcripts to generate short (15–21 base) signature sequences from individual mRNAs (Harbers and Carninci, 2005). Early tag-based sequencing using serial analysis of gene expression (Velculescu et al., 1995) yielded relatively low read depth and required laborious cloning steps. More recently, Illumina's Digital Gene Expression (DGE) platform, upgraded from the previous massively parallel signature sequencing (MPSS) technology (Brenner et al., 2000; Jongeneel et al., 2003; Meyers et al., 2004), can generate, at its current capacity, 90 to 100 million reads per run of an eight-lane flow cell using the Genome Analyzer 2x (GA2x) system (www.illumina.com). Although whole-transcriptome sequencing methods provide information on alternative splicing (Pan et al., 2008; Sultan et al., 2008) and novel expression patterns from intergenic regions (Lister et al., 2008), the nonredundant nature of tag-based profiles would, in theory, allow for greater depth of sequencing per transcript. DGE produces a specific 3′ signature for each mRNA, thereby reducing library saturation from abundant transcripts and enhancing the capacity for rare transcript detection ('t Hoen et al., 2008; Asmann et al., 2009; Morrissy et al., 2009; Babbitt et al., 2010). Likewise, increased read counts per unique transcript would enhance power for statistical analyses in comparing quantitative expression profiles among samples. In addition, DGE data files can be collapsed to a smaller number of unique sequences, thus allowing for less storage requirements, more efficient mapping without the need for high-performance computing, and thus less bioinformatic support. Furthermore, the DGE protocol is strand specific and requires up to 10 times less starting RNA than current whole-transcript, RNA-seq approaches, which is a key advantage when tissue is limiting.

Until recently, an unsequenced genome and lack of adequate gene models and annotations have limited large-scale transcriptome analyses in maize. The maize genome is highly complex, having undergone two successive rounds of duplication (Messing and Dooner, 2006; Wei et al., 2007). Sequencing of the B73 maize reference genome revealed that approximately 81% of the genome could be assigned to homeologous regions (Schnable et al., 2009). In addition, tandemly duplicated gene families occur frequently throughout the genome (Messing and Dooner, 2006; Schnable et al., 2009), and near-identical, paralogous genes (98% or greater identity) are often coexpressed (Emrich et al., 2007b). Such complications have recently been addressed by using sequence-based transcript profiling methods to identify novel genes (Emrich et al., 2007a), resolve the expression of family members and near-identical paralogs (Eveland et al., 2008), and quantify allelic variants (Barbazuk et al., 2007; Guo et al., 2008) in maize.

The shift to functional genomics studies in maize will be dependent on standardized methods for the analysis and assessment of the various sequencing methods with regard to specificity, mapability, depth of coverage, and cost. Despite rapid advances and extensive applications of next-generation sequencing technologies, methods for data analysis have not been well established. In this work, we evaluated the performance of short-read, DGE profiling in cataloging of gene-specific signatures at a particular stage of maize inflorescence development. We present a framework for genome-wide analysis of tag-based expression data in maize and describe a comprehensive pipeline for mapping short sequence reads, accessing gene information from Ensembl (Flicek et al., 2010), and quantifying differences in transcript abundance based on read counts using an open-source statistical package. We also show that analysis of tags mapping independently of known gene models can be used to identify unannotated transcripts, a clear advantage over microarrays. The analyses described here can also be adapted to RNA-seq data sets.

We also tested the effect of perturbing a key developmental pathway in inflorescence architecture using a mutant in the *RAMOSA3* (*RA3*) gene. *RA3* encodes a trehalose-6-phosphate phosphatase (TPP) and functions in regulating the determinacy of axillary meristems (Satoh-Nagasawa et al., 2006). Although *RA3* has been cloned and genetically placed in a pathway controlling meristem determinacy, very little is known about its molecular mechanisms. Here, the quantitative DGE data were used to investigate putative targets of the *RA3* gene. We leveraged functional information available for Arabidopsis and rice (*Oryza sativa*) and resolved the differential expression of transcription factors (TFs) across a wide range of transcript abundance. The significance of this study is 2-fold. First, our results provide a basis for the analysis of short-read, 3′-targeted expression data using the maize B73 genome as a reference. We demonstrate that quantitative differences in transcript abundance can also be detected by DGE with no prior knowledge of the gene space; therefore, it is applicable to species without sequenced genomes. Second, genetic control of branching, especially in the ear where kernels are borne, has clear relevance to crop improvement programs with respect to seed number and harvesting ability.

RESULTS

Library Construction and Sequencing

To generate digital expression signatures for young maize inflorescences, we used the Illumina Genome Analyzer (GA; first phase) technology for massively parallel sequencing by synthesis. In addition, we used a mutant in the *RA3* gene as a developmental perturbation. Immature ears were sampled and hand dissected from field-grown wild-type B73 inbred (Fig. 1A) and *ra3* mutant plants introgressed into a B73 background (Fig. 1B). Ears were size selected for uniformity at a growth stage of 2 mm, where expression of *RA3* is highest (Satoh-Nagasawa et al., 2006) in the wild type (Fig. 1, C and E) and the very first signs of the mutant phenotype were visible as outgrowths of the spikelet pair meristems (Fig. 1, D and F). We represented the wild-type and *ra3* genotypes each with three pools of four to five ears from individual plants. Total RNA was used to construct DGE libraries from each of the six ear samples: three wild-type biological replicates and three *ra3* biological replicates. A single technical replicate of a *ra3* sample was also run.

Briefly, the DGE technology uses a 3′-targeted sequencing strategy to generate a single 20- or 21-base signature tag from the 3′-end of a given transcript. The length of the tag depends on the restriction enzyme used in library construction. We constructed enzyme-specific libraries for each sample using restriction digests with *Dpn*II and *Nla*III. We hypothesized that each enzyme would cleave a given transcript at its 3′-most restriction site and that a dual-enzyme approach would enhance coverage in cases where a restriction site was either absent or within 20 bases of the poly(A) tail. To enable multiplexing of *Dpn*II and *Nla*III libraries, we used a custom sequencing primer that incorporated the restriction site at the 5′-end of each read. The specificity of the restriction site thus allowed for library recognition and sorting of reads sequenced concurrently in a single lane. We sequenced each sample in one lane of an eight-lane Illumina GA flow cell.

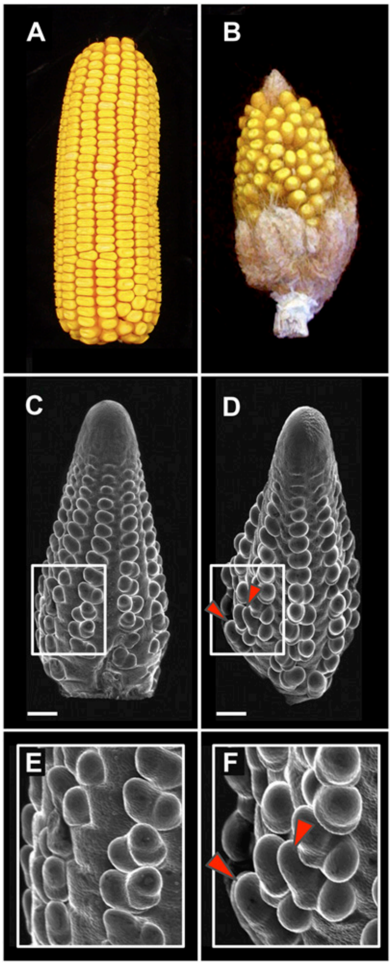

Figure 1. Maize mutants in the *RA3* gene show an increased branching phenotype resulting from a loss of determinacy of basal spikelet pair meristems. A, Mature ear of wild-type B73 maize. B, Mature ear of *ra3* mutant maize. C to F, Scanning electron micrographs at the 2-mm stage show B73 primordia (C and E) and long branches (D and F) just beginning to form at the base of *ra3* mutant ears (red arrowheads). Bars = 200 μm.

In total, approximately 28 million filtered, high-quality reads were sequenced from the seven lanes. Custom Perl scripts were used for adaptor trimming and read parsing. Total reads sequenced per individual sample were $3.9 \pm 1.1 \times 10^6$ (Supplemental Fig. S1A), and about 11% more reads were sequenced from *Nla*III libraries than *Dpn*II. We consolidated reads from all seven lanes into 290,000 and 490,000 unique tags from the *Dpn*II and *Nla*III libraries, respectively. Approximately half of these nonredundant tags were singlets; however, they only represented 1.5% of total reads sequenced. Singlets were removed from further analyses on the basis that they likely represent sequencing errors and there is no statistical support for their presence. Read frequencies of unique signature tags represented by two or more reads (consensus tags) spanned over 4 orders of magnitude (Supplemental Fig. S1B).

Mapping Short Reads to the Maize Reference Genome

We used Vmatch (www.vmatch.de) to map unique consensus sequence tags (total of two or more reads from all libraries) to the maize reference genome (B73 RefGen_v1 [Schnable et al., 2009]). The Vmatch algorithm uses enhanced suffix arrays (Abouelhoda et al., 2002) in which a persistent reference index is created allowing for efficient processing time and reduced space requirements. This method performs effectively with DGE data sets, which are reduced in size and complexity since reads are collapsed to unique tags prior to mapping. Other algorithms for large-scale mapping of short reads can also be used with the condensed DGE data, and we have achieved analogous results using Bowtie (http://bowtie-bio.sourceforge.net).

The short-read mapping pipeline used here included two rounds of mapping to the maize reference genome and associated transcript models (Fig. 2). In phase I, we used a stringent requirement for a complete match of the sequence tag. Here, we allowed a given tag to map perfectly up to three places in the genome. Since we expect that 3′ regions of a given gene tend to be unique, tags mapping to four or more locations were considered repeats and removed from further analyses. In this first round of mapping for *Dpn*II/*Nla*III tags, 45%/54% mapped to a single location, 6%/8% mapped to two or three individual places in the genome, 5%/5% were considered repeats, and 44%/33% did not map (Table I). To determine whether a portion of these unmapped tags covered splice junctions, we used the transcript models associated with the maize reference sequence (www.maizesequence.org) as a persistent Vmatch index. We recovered an additional 1.7%/4.8% (*Dpn*II/*Nla*III) of total tags that mapped to a single location in the transcriptome and 0.4%/0.9% that matched two or three transcripts.

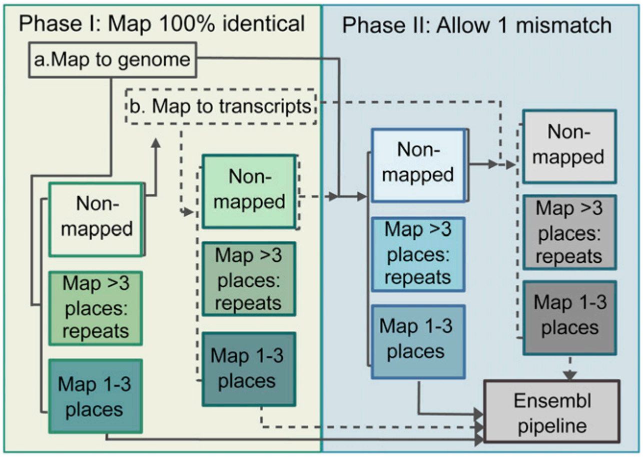

Figure 2. Bioinformatics pipeline used to map unique DGE tags to the maize reference genome. In phase I, only perfect matches were allowed. Tags that mapped uniquely up to three places in the genome or associated cDNA models were used in downstream analyses to extract Ensembl-based gene information. Tags that did not map during phase I were subject to an additional round of mapping (phase II), which allowed for one mismatch along the length of the sequence tag. All tags that mapped to repeat regions of the genome (more than three unique places) were removed from further analyses.

In phase II, a second round of mapping used the remaining 42%/31% (*Dpn*II/*Nla*III) of tags that did not map completely to the genome or associated transcript models. Here, we allowed for one mismatch to maximize the recovery of signatures that did not map due to sequencing errors in the reference or polymorphisms retained after introgression of the *ra3* mutant. While the one-mismatch tags tended to be distributed uniformly across the genome, we did observe an enrichment of tags flanking the *RA3* locus on chromosome 7 that were sequenced exclusively from the *ra3* samples and represented by at least 10 reads (Supplemental Fig. S2). Although these made up only 0.2% of all unique mapped tags, they could potentially be used to resolve areas of variation associated with the introgression. After two rounds of mapping, only 14%/13% (*Dpn*II/*Nla*III) of all unique tags did not map to the reference maize genome sequence or associated transcripts (Table I). These nonmapped tags most likely represent regions where the reference sequence is incomplete or varies between B73 and the original *ra3* mutant line. Only 0.02% of nonmapped tags matched maize chloroplast or mitochondrial genome sequences.

Extracting Gene Information for Mapped Tags

In the next stage of our analysis pipeline, we used the mapping coordinates for tags that matched one to three unique places in the genome and extracted the corresponding gene information. Custom scripts used the Ensembl Perl Application Programming Interface (http://uswest.ensembl.org/info/docs/api) to associate mapped tags with a “working” gene set of 108,745 gene models including evidence-based (86%) and ab initio (14%) predictions (gene build 4a.53; maizesequence.org). The working gene set is a broader, less conservative set of gene models, which include a smaller set of high-confidence, “filtered” gene models (gene build 4a.53; maizesequence.org). We anticipated that by using these working genes, our analysis would not be restricted to well-characterized genes and thus enhance the potential for gene discovery. In order to maximize the inclusion of unannotated untranslated regions (UTRs), the predicted gene space was extended 300 bases on either end. In total, we identified 37,117 working genes (including 22,500 filtered genes) that were associated with at least two read counts in *Dpn*II and/or *Nla*III libraries. Of these, 9% were *Dpn*II specific while 21% were found only in the *Nla*III data set. We then

Table I. *Summary statistics from mapping unique sequence tags (represented by two or more reads) to the maize B73 reference genome*

Mapping Class	No. of Unique Tags (% of Total)[a]			
	Phase I		Phase II	
	Genome	Transcripts Only	Genome	Transcripts Only
*Dpn*II libraries				
Total unique tags	134,656	58,746	57,503	19,431
Unique match	60,418 (44.9)	995 (1.7)	20,525 (35.7)	111 (0.6)
Two to three matches	8,668 (6.4)	222 (0.4)	11,084 (19.3)	59 (0.3)
Repeats	6,824 (5.1)	243 (0.04)	6,463 (11.2)	4 (0.2)
Nonmapped	58,746 (43.6)	57,503 (97.9)	19,431 (33.8)	19,257 (99.1)
*Nla*III libraries				
Total unique tags	237,005	78,662	74,034	30,597
Unique match	128,332 (54.2)	3,794 (4.8)	26,708 (36.1)	289 (0.9)
Two to three matches	18,775 (7.9)	702 (0.9)	11,583 (15.7)	137 (0.4)
Repeats	11,236 (4.7)	132 (0.2)	5,146 (7.0)	21 (0.1)
Nonmapped	78,662 (33.2)	74,034 (94.1)	30,597 (41.3)	30,150 (98.5)

[a](% of Total) represents the portion of tags mapped from total tags subjected to each stage of mapping.

compared total read frequencies to the percentage of mapped tags associated with working gene models, nongenic space, and repetitive regions and to nonmapped tags (Fig. 3). The proportion of tags that mapped to a working gene model increased from approximately 40% at low read counts (two to 10 reads) to 80% at 1,000 reads or more. Overall, there were more of the low-frequency tags sequenced from the *Nla*III libraries (Supplemental Fig. S1B), and a higher percentage of these mapped to genic regions compared with those from *Dpn*II libraries. Tags that mapped to repetitive regions represented about 6% of the total in each frequency class (Fig. 3), and the majority of nonmapped tags were low copy (50% and 40% for *Dpn*II and *Nla*III libraries, respectively). We also observed a small proportion of tags in each read count class that mapped to regions of the genome where no gene model was called. When mapping from phase II (one mismatch allowed) was considered, an additional 2,936 working genes were identified; however, most of these were recovered from low-frequency tags (Supplemental Fig. S3).

Since the number of genes recovered (37,117) was only 15% of the total number of mapped tags, multiple signature tags were likely associated with a single gene. Therefore, we combined consensus sequence tags that mapped to a given gene model to obtain a cumulative read count (Fig. 4A). Only 36% and 24% of the genes identified in *Dpn*II and *Nla*III libraries, respectively, were associated with a single signature tag. We expect that multiple tags are due to incomplete restriction enzyme digests during library preparation; however, a portion may represent alternate splice isoforms or polyadenylation variants. The number of tags per gene was distributed over a wider range in the *Nla*III data set, with some genes associated with more than 20 tags. This distribution was also observed with an independent maize *Nla*III data set constructed and sequenced in a different laboratory (P. Bommert, unpublished data), suggesting that this was not due to technical errors during library construction.

Although multiple tags mapped to a given gene, we expected that the most abundant tags would be those associated with the 3′-most restriction site for each gene. To test this, we used a set of 36,394 full-length maize cDNAs (fl-cDNAs; Alexandrov et al., 2009) as a "golden" reference set of transcript models with

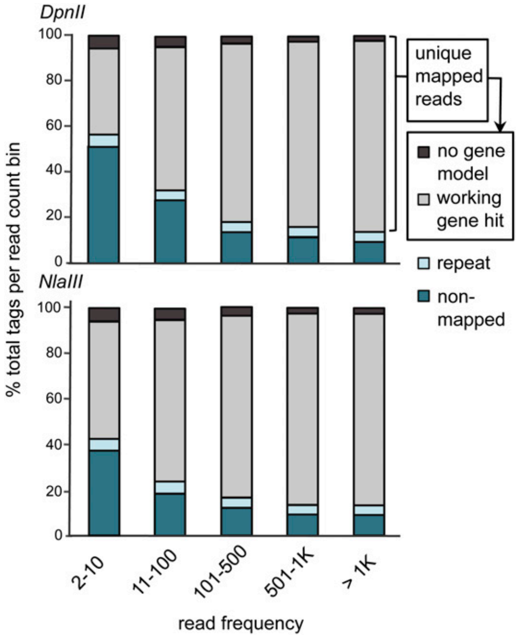

Figure 3. Mapping results for unique consensus tags (represented by two or more reads) and distribution by total read count. Coordinates of mapped tags from phase I were used to associate sequence reads with a working set of maize gene models (gene build 4a.53; maizesequence.org). Tags were classified as mapping to working gene models or nongenic regions, repeats, or not mapping at all. Results are displayed as consensus tags, grouped by total read count across all samples. In both *Dpn*II (top) and *Nla*III (bottom) data sets, nonmapped tags tended to be low copy (two to 10 reads), while the largest portion of tags at read frequencies greater than 100 tended to be associated with gene models.

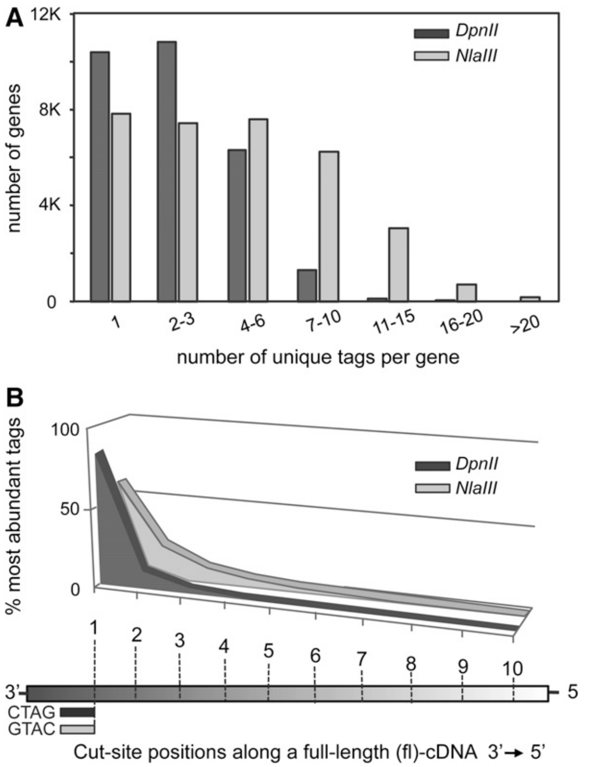

Figure 4. Distribution of genes associated with multiple unique tags and relative frequencies of tags at 3′ enzyme cut sites. A, The total number of maize working genes identified in *Dpn*II and *Nla*III data sets were grouped according to number of unique tags that map to them. B, Digests were simulated in silico with *Dpn*II and *Nla*III on a golden set of fl-cDNA models. Positions of predicted 20- to 21-nucleotide fragments that matched unique DGE tags were used to estimate frequency (shown as percentage) of the most abundant tag per fl-cDNA generated by each possible site from the 3′-end.

mostly complete 3′-UTRs. We simulated *Dpn*II and *Nla*III digests in silico and matched the predicted DGE tags from the resulting fragments to our consensus set of unique sequence tags. In total, 67% and 74% of the fl-cDNAs were detected in *Dpn*II and *Nla*III libraries, respectively. For each fl-cDNA, associated DGE tags were ranked based on position starting from the 3′-most and extending to the 10th possible restriction site. The percentage of tags representing the most abundant signature was quantified (Fig. 4B). Based on these data, the canonical, 3′-most tag was the most abundant for 85% of fl-cDNAs identified by *Dpn*II and 60% with *Nla*III.

Since strand information is retained during construction of the DGE libraries, we were also able to distinguish sense and antisense transcripts. Of the 37,117 working genes identified in this experiment, approximately 49% showed evidence for transcription in both orientations, while about 9% showed antisense expression alone. There were 2,346 working genes represented by at least 10 reads in both sense and antisense orientations for wild-type and/or *ra3* samples. We analyzed sense-antisense (S-AS) pairs based on methods by Morrissy et al. (2009). Here, the ratio of antisense relative to sense transcript abundance was calculated for each gene and used to determine if S-AS transcript ratio (S:AS) was altered in the *ra3* mutant. Ratio changes ranged from +5.9 to −6.2, where positive values indicated a decreased S:AS in the *ra3* mutant compared with the wild-type. In *Dpn*II and *Nla*III data sets, 86 and 170 genes, respectively, had S:AS changes that were at least 2-fold in *ra3* mutants compared with the wild type (Supplemental Fig. S4; Supplemental Table S1).

Determining Differential Gene Expression

To compare gene expression profiles for the wild type and mutant, we first used cumulative counts of all consensus tags mapping in sense to a given working gene model (Supplemental Fig. S5). Initially, only tags that mapped completely to the genome and/or transcript models (phase I) were used, and data from *Dpn*II- and *Nla*III-generated data sets were analyzed separately. Since we included tags that mapped to two or three locations in the genome, we applied a scoring system to normalize read count across mapping locations (see "Materials and Methods"). We verified that read distribution across lanes was not skewed due to saturation of very-high-frequency sequences by plotting raw read counts for the most abundant tags encountered in *Dpn*II and *Nla*III libraries (Supplemental Fig. S6).

We used edgeR (empirical analysis of digital gene expression in R; Robinson and Smyth, 2008; Robinson et al., 2010), a software package available from Bioconductor (Gentleman et al., 2004), to normalize for tag distribution per library and determine significance values for differentially expressed genes. The edgeR algorithm uses an empirical Bayes analysis to improve power in small sample sizes (Robinson and Smyth 2007, 2008; Robinson et al., 2010). This accounts for biological and technical variation and has been implemented for tag-based data sets where small numbers of replicates are tested and SE values disperse farther from the mean at low versus high levels of expression (Robinson and Smyth 2008; Babbitt et al., 2010). Based on a cutoff of at least nine reads per gene, our statistical analyses included 20,250 and 22,130 genes for *Dpn*II and *Nla*III data sets, respectively (Supplemental Data Set S1). From these, we identified 660 and 303 differentially expressed genes, respectively, with false discovery rate-corrected P values less than 0.05 (Fig. 5A; Supplemental Fig. S7). A smaller number of significantly different genes identified in the *Nla*III data set is likely due to more variation observed among *Nla*III samples and small sample sizes. Among all differentially expressed genes ($P < 0.05$), 629 were up-regulated in the *ra3* mutant while 249 were down-regulated. Many of the genes showing the most significant differences in expression have not been characterized in maize or in related species (Supplemental Table S2).

Although 74% of the total expressed gene set used in the edgeR analysis was present in both *Dpn*II and

*Nla*III data sets, only 9% of these genes had significant *P* values (<0.05) in each of the two data sets (Fig. 5A; Supplemental Table S3). Although the remaining 65% (675 genes) had significant *P* values exclusively in *Dpn*II or *Nla*III data sets, the majority of these (97.4%) showed the same trend (i.e. either up- or down-regulated in the *ra3* mutant) in both data sets (Fig. 5B). As expected, a larger proportion of genes that were highly significant ($P < 1.0e^{-08}$) in one of the two data sets also showed significant expression differences in the other (Fig. 5A). We used quantitative real-time PCR (Q-PCR) with independently collected RNA samples to test whether expression differences were reproducible for a subset of genes. Significant differences were validated for 80% of the genes tested using Q-PCR (Fig. 5C).

Expression Profiling with No Prior Knowledge

To determine differences in transcript abundance independently of prior information on gene models, we carried out statistical testing using edgeR for each mapped tag separately (Supplemental Data Set S2). In this analysis, we used tags that mapped completely to

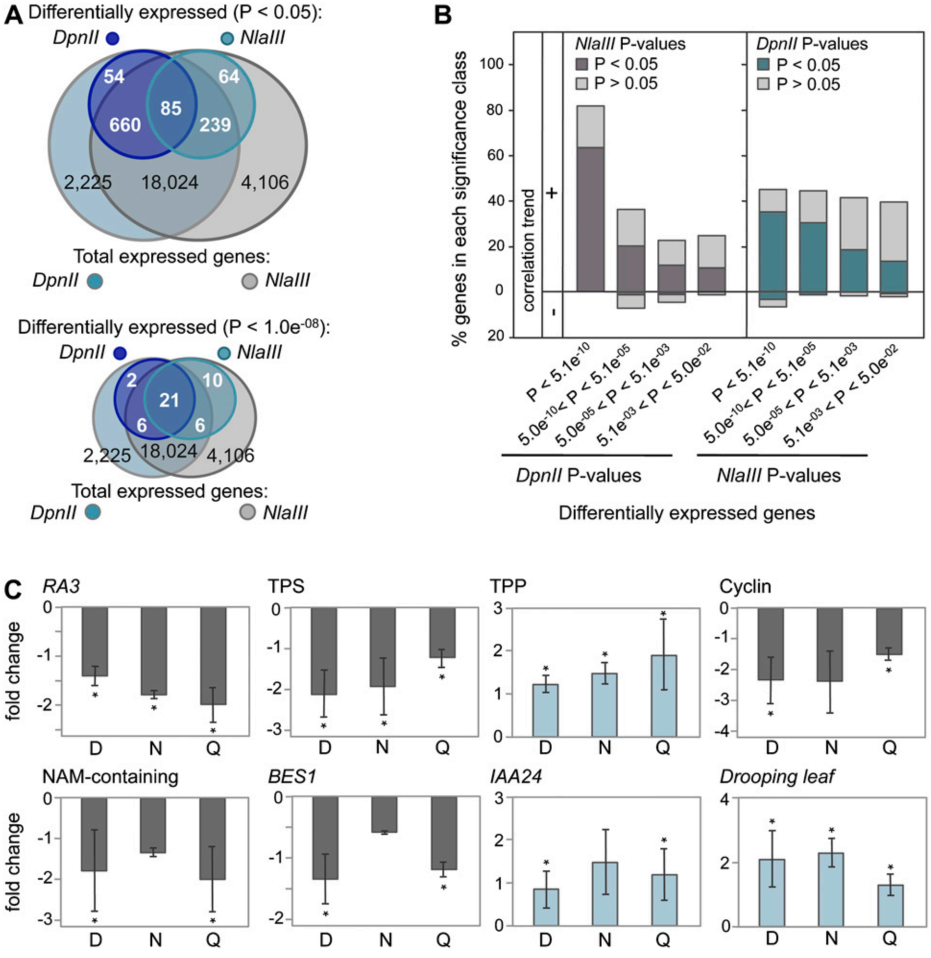

Figure 5. Differentially expressed genes identified in the *Dpn*II and *Nla*III data sets. A, Venn diagrams show total number of expressed genes (represented by nine or more reads) analyzed from *Dpn*II and *Nla*III data sets and the portion that were differentially expressed between the wild type and *ra3* with corrected *P* values of $P < 0.05$ (top) and $P < 1.0e^{-08}$ (bottom). Genes identified as differentially expressed in both data sets tended to have smaller *P* values. B, Expression trends (up- or down-regulated in *ra3*) were compared for genes that were identified in both data sets but had significant *P* values ($P < 0.05$) in only *Dpn*II (left panel) or *Nla*III (right panel). For each data set, differentially expressed genes were divided into four classes based on significance level. Within each class, the percentage of genes that were differentially expressed in both data sets are shown along with whether they shared common (+) or had opposite (−) trends. Genes with significant *P* values in either *Dpn*II or *Nla*III data sets, but a *P* value of 1.0 (no change in expression) in the other, were not included. C, Differential expression was validated by Q-PCR for a subset of genes. Relative fold changes are shown for *Dpn*II (D) and *Nla*III (N) data sets and for Q-PCR data (Q). Some genes were differentially expressed in only one DGE data set (asterisks) but showed common trends (significant *P* not always consistent with fold change). TPS, Trehalose phosphate synthase; NAM, NO APICAL MERISTEM; BES1, BRI1-EMS-Suppressor; IAA24, auxin/indole-3-acetic acid family protein 24.

the genome and were independently represented by at least nine reads, rather than cumulative counts per gene. We identified 364 unique tags from *Dpn*II libraries and 294 tags from *Nla*III libraries that showed significant differences in frequency between wild-type and *ra3* mutant samples. These tags were grouped into four hierarchical classes based on level of significance (where class I included tags with most significant *P* values and class IV included tags with least significant *P* values; all classes were <0.05). Their individual read frequencies were then plotted according to tag rank order (Fig. 6). Mapping the tags back to their respective gene models resulted in comparable sets of differentially expressed genes from *Dpn*II and *Nla*III data sets, as observed with cumulative counts (data not shown). Results from this analysis showed that although more counts per tag provided greater power for statistical testing, we could also capture significant differences for low-frequency tags with read counts just above 10.

These data indicate that differential gene expression can be analyzed without prior knowledge of gene models and that significant differences can be detected for low-abundance transcripts. Therefore, with ongoing upgrades in throughput for Illumina and other sequencing technologies, the ability to multiplex DGE samples can dramatically improve cost and time efficiency. In addition, these analyses provide a more unbiased approach and an advantage over microarray design, since existing gene models may not be complete or capture all transcript variants. We further tested whether tags that mapped to unannotated regions of the genome were of potential biological significance. Of those individual mapped tags with significant differences in frequency between the wild type and mutant (Fig. 6), 37 *Dpn*II and 25 *Nla*III unique tags did not associate with working gene models. We used mapping coordinates of these 62 unique tags to cluster those that mapped within 5 kb of each other. This identified four unique regions where these differentially expressed tags clustered (Supplemental Fig. S8). Among these, a 554-nucleotide region on chromosome 8 and a 488-nucleotide region on chromosome 4 were associated with tags expressed only in *ra3* samples. Although no gene model has been called in either of these regions, available RNA-seq data for maize (maizesequence.org; Schnable et al., 2009) provide additional evidence for expression. A 57-nucleotide region on chromosome 5 and a 531-nucleotide region on chromosome 7 were each associated with two differentially expressed tags and appear to be unannotated 3′-UTRs to adjoining genes (Supplemental Fig. S8). This clustering method could also be done with all mapped tags to obtain cumulative read counts in the absence of predicted gene models.

Resolving TFs across a Wide Range of Abundances

A primary objective was to identify genes that encode TFs and to determine potential ranges for their detection. To test this, we used information from the Ensembl Compara gene trees (Vilella et al., 2009) at maizesequence.org and gramene.org (Liang et al., 2008) to retrieve putative orthologs of maize genes in our expressed set. We then queried known Arabidopsis TFs in the Database of Arabidopsis Transcription Factors (http://datf.cbi.pku.edu.cn/) and between both *Dpn*II and *Nla*III data sets identified 479 maize genes with sequence similarities to Arabidopsis TFs. Quantitative analysis of their expression profiles indicated that TFs are expressed over a wide range of transcript abundances, spanning over 4 orders of magnitude in young inflorescence tissue (Fig. 7; Supplemental Data Set S3). Of the 479 putative TFs, 27 were differentially expressed ($P < 0.05$) based on

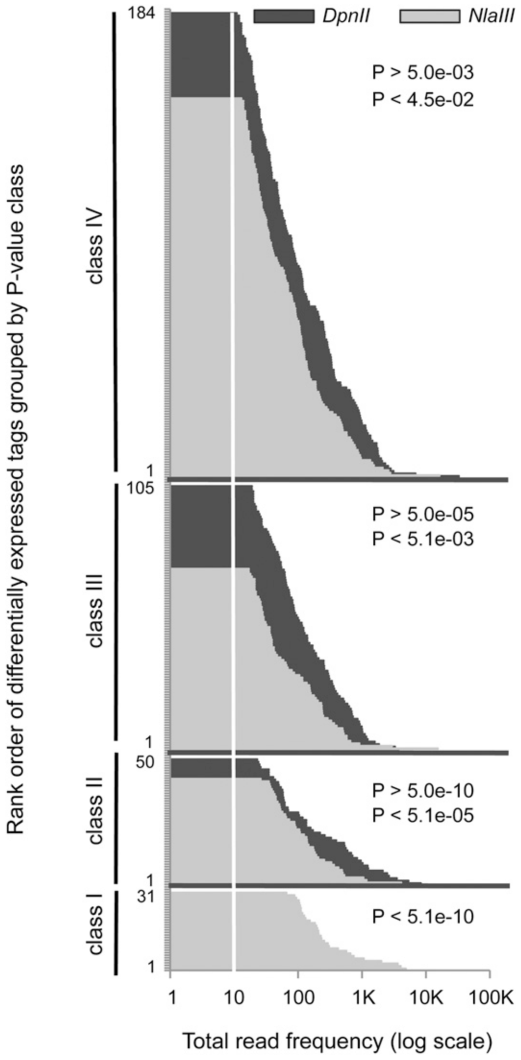

Figure 6. Distribution of read frequencies for unique consensus tags that showed differential expression between the wild type and *ra3* mutant. Total read counts for each tag are plotted on the *x* axis (log scale). Tags are grouped into four significance classes based on their *P* value and are plotted by their respective rank order in each class (*y* axis). A read count of 10 was used as a cutoff for analysis of differential expression (marked by the white line).

edgeR statistical analysis. We further interrogated the remaining differentially expressed gene set for additional TFs using gene ontologies, InterPro domains, and known maize annotations.

Together, these analyses identified a total of 75 putative TFs, which were differentially expressed over a wide range of abundances (Fig. 7; Supplemental Table S4). A subset presented in Table II includes members of TF families based on InterPro signatures (www.ebi.ac.uk/interpro) associated with functions in development and meristem maintenance or identity (NAC, YABBY, GRAS), while others have roles in hormone-mediated signaling by auxin (AUX/IAA), brassinosteroids (BES, BIM), or ethylene (AP2/ERF). Among the 75 differentially expressed TFs, nine were characterized as AP2/ERF family proteins (PF00847) based on InterPro and PFam classifications (Fig. 7).

DISCUSSION

In this study, we developed and tested a framework for analysis of short-read, sequence-based expression profiles using Illumina's DGE technology and the first assembled maize reference genome (B73 RefGen_v1 [Schnable et al., 2009]). Our results demonstrate that deep sequencing of 20- to 21-nucleotide DGE tags can be used to successfully resolve genome-wide expression profiles in maize and detect differences in transcript abundance over a broad dynamic range. In our test case, which compared three pooled samples each of wild-type and *ra3* ears, we identified 37,117 expressed maize working genes (including 22,700 high-confidence filtered genes) from six DGE libraries, each sequenced in a single lane of an Illumina GA flow cell. Of these, 67% were represented by sense transcripts with nine or more reads and were used to test for significant differences in transcript abundance between the wild type and *ra3* mutants. Results from these expression analyses provide testable hypotheses for resolving regulatory and biochemical processes contributing to maize inflorescence architecture via the *RA* pathway.

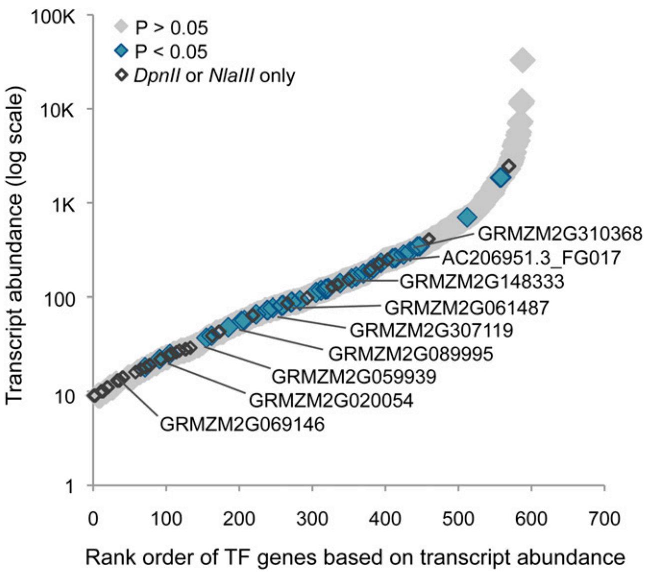

Figure 7. Quantitative transcript profiles for putative TFs expressed in 2-mm ears. These include 479 genes with sequence similarity to known TF genes in Arabidopsis. Also shown are 75 differentially expressed genes ($P < 0.05$) encoding possible TFs identified by comparative analyses with Arabidopsis, GO assignments, and InterPro signatures. Transcript abundances are plotted on a log scale based on read frequencies for each TF and distributed along the *x* axis according to rank order. TFs found exclusively in *Dpn*II or *Nla*III data sets are highlighted. Nine AP2/ERF family members (noted here by maize gene identifiers) were identified among differentially expressed TFs.

Evaluation of DGE-Based Analysis with the Maize Reference Genome

One objective of this work was to evaluate the performance of DGE as a high-throughput method for genome-wide transcript profiling in maize. As deep sequencing technologies continue to develop, their ability and/or efficiency for addressing specific research questions will be based on thresholds that exist for read length, depth of coverage, sequence specificity, and cost. One limitation of the Illumina tag-based DGE platform is the short read length (20–21 nucleotides); however, its exceptional throughput at a lower cost promotes DGE as a candidate for use in large-scale expression profiling experiments. We speculated that the availability of a sequenced maize genome would improve our ability to map the short DGE tags. From the conservative mapping phase (phase I) of our DGE analysis pipeline, we were able to map 51% of unique sequence tags to a single location in the maize genome. Although 35% of unique tags did not map in phase I, our results are consistent with those of 't Hoen et al. (2008), who showed that 41% of DGE tags could not be mapped in mouse. By performing an additional round of mapping (phase II, one mismatch allowed), we were able to recover 63% of these nonmapped tags. The remaining 14% were most likely due to incomplete regions of the reference sequence and variation due to introgression.

Other studies using DGE ('t Hoen et al., 2008; Morrissy et al., 2009) and MPSS (Meyers et al., 2004; Nobuta et al., 2007) included only those tags that mapped unambiguously to the reference sequence in their analyses. One feature of the Vmatch mapping software used here is the ability to view all redundant matches to the genome. In this work, tags that mapped perfectly to two or three individual loci represented 9% of the total. These were included in our analyses with the expectation of recovering information on paralogous genes. For example, for a set of AP2/ERF family members, we could successfully resolve closely related paralogs in the maize genome and quantify their unique transcript profiles. Our confidence for detecting paralogous loci was improved by combining

Table II. *Differentially expressed maize genes were identified as putative TFs*

*, Q-PCR validations were done for a subset of genes.

Maize Gene ID	Annotation[a]	TF Family	Read Frequency[b]		LogFC[c]	*P* Value[d]
			Wild Type	*ra3*		
Up-regulated TF genes in *ra3* mutants						
GRMZM2G307119	Branched silkless1	AP2/ERF	0 ± 0	18 ± 12.8	$3.38e^{+01}$	$4.06e^{-09}$
			0 ± 0	19 ± 10.6	$3.36e^{+01}$	$1.28e^{-10}$
GRMZM2G088309	Drooping leaf*	YABBY	1.4 ± 0.6	14 ± 8.3	3.52	$8.43e^{-05}$
			6.0 ± 1.2	55 ± 13	3.17	$1.19e^{-12}$
GRMZM2G127379	NAM containing	NAC	1.7 ± 0.8	22 ± 9.2	3.67	$1.52e^{-07}$
GRMZM2G017606	SHI	SHI	11 ± 0.7	43 ± 14	2.05	$8.42e^{-05}$
GRMZM2G061487	DRE binding factor 1	AP2/ERF	8.3 ± 2.7	22 ± 6.9	1.54	$3.07e^{-02}$
			10 ± 1.5	36 ± 7.8	1.83	$4.11e^{-04}$
GRMZM2G055243	KNOX class 2 protein	KNOX	4.3 ± 0.3	21 ± 12	2.39	$4.73e^{-04}$
GRMZM2G089995	Ethylene responsive	AP2/ERF	2.3 ± 0.4	14 ± 8.1	2.67	$7.59e^{-04}$
GRMZM2G310368	Ethylene responsive	AP2/ERF	34 ± 9.5	81 ± 14	1.33	$8.95e^{-04}$
GRMZM2G171852	Uncharacterized	C2C2-Dof	26 ± 1.2	60 ± 3.8	1.37	$5.63e^{-03}$
GRMZM2G078077	TCP domain protein	TCP	9.0 ± 0.5	27 ± 10	1.73	$5.83e^{-03}$
			4.3 ± 1.2	16 ± 3.5	2.12	$2.46e^{-03}$
GRMZM2G003927	Ramosa1	Znf-C2H2	107 ± 10	237 ± 31	1.27	$2.71e^{-03}$
			40 ± 13	97 ± 24	1.27	$1.59e^{-03}$
GRMZM2G020054	Uncharacterized	AP2/ERF	0.4 ± 0.3	7.0 ± 3.2	4.52	$2.34e^{-03}$
GRMZM2G014653	NAC protein 48	NAC	8.7 ± 3.3	29 ± 8	1.70	$3.22e^{-03}$
GRMZM2G132367	HDZipI-1	HD-Zip	7.7 ± 2.0	23 ± 5.7	1.58	$1.67e^{-02}$
GRMZM2G115357	IAA24*	AUX/IAA	12 ± 3.2	30 ± 4.1	1.48	$1.29e^{-02}$
GRMZM2G130149	MYB59	R2R3-MYB	14 ± 3.2	34 ± 9.2	1.29	$3.31e^{-02}$
Down-regulated TF genes in *ra3* mutants						
GRMZM2G088242	HSFB4	HSF	198 ± 30	36 ± 11	2.36	$9.51e^{-09}$
			17 ± 1.8	3 ± 2.8	2.80	$8.23e^{-04}$
GRMZM2G102514	BES1/BZR1 protein*	BES	499 ± 33	123 ± 43	1.91	$1.85e^{-06}$
GRMZM2G171468	Uncharacterized	MYB	88 ± 8.1	31 ± 17	1.35	$6.52e^{-03}$
			33 ± 6.3	6 ± 1.4	2.48	$2.22e^{-05}$
GRMZM2G054277	NAM containing*	NAC	17 ± 3.4	3 ± 1.4	2.58	$9.84e^{-04}$
GRMZM2G148333	Ethylene responsive	AP2/ERF	30 ± 3.9	5.7 ± 4.5	1.92	$1.78e^{-03}$
GRMZM2G116658	Outer cell layer3	HOX	18 ± 1.7	3.7 ± 1.9	2.20	$1.39e^{-03}$
GRMZM2G051955	ZF-homeobox protein	ZF-HD	21 ± 3.8	5.5 ± 4.9	1.93	$1.03e^{-02}$
GRMZM2G089501	BIM2	BIM	15 ± 5.8	3.3 ± 1.3	2.07	$1.53e^{-02}$
GRMZM2G172657	Uncharacterized	GRAS	70 ± 2.5	32 ± 13	1.13	$3.19e^{-02}$

[a] Annotations are based on Ensembl gene descriptions at maizesequence.org, gene build 4a.53. [b] Read frequency is average read count ± SE (reads per million) for the three biological replicates in wild-type and *ra3* samples. [c] Log fold changes from edgeR analysis of differential gene expression. [d] Corrected *P* values (false discovery rate of 5%) from edgeR analysis of differential gene expression. If $P < 0.05$ in both enzyme libraries, values for *Dpn*II and *Nla*III are shown, respectively.

tags from *Dpn*II and *Nla*III data sets that associated with a given gene model (Supplemental Table S5).

Similar to observations in this work, previous studies with tag-based profiling methods using either the *Dpn*II (Meyers et al., 2004) or the *Nla*III ('t Hoen et al., 2008; Babbitt et al., 2010) enzymes also showed that multiple tags tended to associate with a single gene model. Although these have been suggested to represent polyadenylation variants, we expect, based on the high frequencies and wide distribution of tags, that many are due to incomplete restriction enzyme digestion during library preparation and/or enzyme biases for specific cut sites. In support of this, analysis of the biological replicates in our study showed that read frequencies of individual tags varied among replicates to a larger degree than the cumulative counts for a given gene (Supplemental Fig. S5). Profiling with cumulative counts also allowed for the inclusion of more genes in our statistical analysis, since a cutoff of nine reads was imposed as a threshold for detection. In addition, our analysis using the full-length cDNA models indicated that the 3′-most tag was not always the most abundant signature for a gene. This was more commonly observed with the *Nla*III enzyme, which further suggests that multiple signatures per gene were not due to variation in transcript structure.

Based on these observations, we used a cumulative count of tags mapping to a given gene, as opposed to previous DGE and MPSS studies that used a predicted 3′-tag database for detecting individual transcripts (Meyers et al., 2004; 't Hoen et al., 2008; Morrissy et al., 2009). Relying on the latter method in a complex genome such as maize, which is in the early stages of annotation, could result in a loss of informative expression data. Accordingly, recent work using RNA-seq data to improve gene models in human identified extensive unannotated UTRs (Pickrell et al., 2010). In our analysis, although we expect to lose the resolution of alternate transcripts, such as 3′-RNA processing variants, combining tags on a per-gene basis, including tags mapping within 300 bases of a predicted UTR, provides more comprehensive expression analyses for poorly annotated genes. Combining DGE data sets with long-read or paired-end RNA-seq approaches would likely improve confidence for identifying alternatively spliced transcripts.

Analysis of Enzyme-Specific Data Sets

One ambiguity of sequencing methods that use restriction enzymes in library construction is the potential bias and/or efficiency of enzymes for specific sequences (Siddiqui et al., 2006). Our study represented the first direct comparison, to our knowledge, of two different enzymes across a set of biological replicates. On average, there were 1.7 times more unique tags sequenced from *Nla*III libraries compared with *Dpn*II. This may be expected, since *Nla*III cuts approximately 1.5 million more times in the maize genome, and approximately 132,000 more times in the transcriptome, than *Dpn*II. In addition, *Nla*III generated more unique tags per gene model overall (in some cases up to 20 individual tags per transcript) and more primary signature tags from 3′-most restriction sites when compared with *Dpn*II. These data suggest that processing of alternate cut sites is not random between the two enzymes and that DGE libraries generated from a single enzyme may be prone to biases.

Based on observations from unrelated experiments that *Nla*III produces more tags per gene (P. Bommert and M. Regulski, unpublished data), it is possible that noise from partial restriction enzyme digestions can skew expression profiles and dilute biologically relevant information. Other factors, such as a slight G+C bias in the *Nla*III library (Supplemental Fig. S9), amplification biases prior to sequencing, or the general instability of the *Nla*III enzyme, may contribute to the limited overlap of differentially expressed genes between *Dpn*II and *Nla*III data sets. This could further be explained by the small sample sizes typically used in sequence-based profiling experiments, which are subject to false positives, and more variation among *Nla*III samples, presumably due to technical bias during library construction. However, in this study, our approach to analyze *Dpn*II and *Nla*III data sets independently enhanced the power to detect highly significant changes in gene expression while decreasing false positives.

Other sequence-based expression studies have found that analyses of low-copy transcripts were often unreliable, even in the absence of enzyme-based library construction (Marioni et al., 2008; Mortazavi et al., 2008; Fahlgren et al., 2009). Here, although we observed less variation between replicates for highly expressed genes, differences in expression trends between the two data sets did not seem to correlate with total read count (data not shown). Overall, despite variations between the enzyme-specific data sets, we showed that the dual-enzyme approach provided expression data on a more complete panel of genes as well as validation for a high-confidence set of genes identified as significant from both data sets. The latter was especially true for genes with highly significant expression differences between test groups. Furthermore, we showed that expression profiles identified in both DGE data sets could be experimentally validated by Q-PCR.

Applications for Functional Analyses with DGE Data

After identifying differentially expressed genes, the next step is to ask whether these genes reveal functionally relevant information. However, most genes with significant differences in transcript abundance were largely uncharacterized in maize or closely related plant species. This is due, in part, to the fact that functional ontologies used to classify genes (i.e. Gene Ontogeny [GO] and Pfam) are primarily based on bacterial and animal models, and many plant-specific genes have not been functionally annotated. Of the total expressed genes that were used in statistical testing (22,267 and 24,997 from *Dpn*II and *Nla*III data sets, respectively), only 48% were associated with GO terms. Consequently, this impacted our ability to resolve significant enrichment for gene ontologies in the DGE data sets. However, we found overrepresentation of intercellular (Cellular Compartment, GO:5622; $P = 1.94e^{-04}$) and RNA binding (Molecular Process, GO:3723; $P = 1.84e^{-02}$) in the differentially expressed gene set. These results are consistent with the predicted roles for *RA3* in cell-to-cell signaling and gene regulation (Satoh-Nagasawa et al., 2006). We anticipate that as expression data sets for maize are generated, functional annotations will improve through the integration of metadata and the curation of coexpressed genes and pathways (Horan et al., 2008).

Although a large proportion of genes that showed significant transcriptional changes have not been characterized in maize, we were able to leverage known functional information from Arabidopsis and rice to identify putative classes of metabolic and regulatory genes. For example, rice genes have been associated with biochemical pathways, and we used the ricecyc pathway tool (www.gramene.org; Liang et al., 2008) to determine whether differentially expressed genes were assigned to common metabolic pathways. For this, we used a significance threshold of $P < 0.08$ for differential expression in order to increase the coverage of pathways and identified 781 putative rice orthologs using the Ensembl Compara gene trees (maizesequence.org, gramene.org; Vilella et al., 2009). Of these, 67 (54 up-regulated and 13 down-regulated in *ra3* mutants) could be mapped onto 97 specific metabolic pathways (Supplemental Table S6).

The *RA3* gene is expressed in a narrow band subtending the maize spikelet pair meristems during early inflorescence development (Satoh-Nagasawa et al., 2006). As a TPP, it is possible that a mobile signal, such as a sugar, could be mediating *RA3*'s control of axillary meristem cell fate (Rolland et al., 2006). We observed that many of the differentially expressed genes that could be mapped onto metabolic pathways were associated with primary carbohydrate biosynthesis and degradation, respiration, and energy production as well as redox and nitrogen cycling processes (Supplemental Table S6). As expected, trehalose biosynthesis was represented in the differentially expressed set. Expression of the *RA3* gene was significantly down-regulated in the mutant, as was a trehalose phosphate synthase (GRMZM2G077659). In contrast, an uncharacterized gene in maize (GRMZM2G151044) with sequence similarity to a TPP in Arabidopsis was up-regulated in the mutant, possibly as compensation for reduced *RA3* levels (Fig. 5C; Supplemental Fig. S10).

Disruption of trehalose biosynthesis in the *ra3* mutant could have global affects on the sugar status of the cell due to altered Glu-6-P and trehalose-6-phosphate levels. We found that pathways for transient starch degradation (Smith et al., 2005) and downstream reactions that utilize hexoses as substrates (i.e. Glu-6-P) were also represented in the differentially expressed set. These included glycolysis and the oxidative pentose phosphate pathway, both of which generate reducing power in the form of NAD(P)H. The intermediates generated by the reaction of these enzymes represent potential signals that report the sugar, redox, or adenylate status of the cell (Supplemental Fig. S10). In addition, a number of genes that encode enzymes with oxidoreductase activities were differentially expressed and were primarily up-regulated in the mutant. These activities also have the potential to generate signals. For example, previous work has implicated the trehalose intermediate trehalose-6-phosphate as mediating the redox regulation of a key starch biosynthetic enzyme (Kolbe et al., 2005).

Aside from a potential metabolic role, it has also been proposed that *RA3* may have a transcriptional regulatory role due to its mutant phenotype, which is shared with two known TFs, *RA1* and *RA2*. Dual biochemical and transcriptional activity is reminiscent of other sugar-responsive metabolic genes, such as *HEXOKINASE*, which has been shown to function in a transcriptional complex to regulate gene expression (Cho et al., 2006). Of the genes that showed significant differences in expression, 72% were up-regulated in the *ra3* mutant, suggesting that *RA3* could act primarily by repressing transcription. Our analysis of TFs revealed putative candidates for downstream analyses to determine if they are direct or indirect targets of *RA3*. Among these, *RA1* (GRMZM2G003927; Vollbrecht et al., 2005) has been genetically placed in the same developmental pathway as *RA3* (Satoh-Nagasawa et al., 2006), and *BRANCHED SILKLESS1* (*BD1*; GRMZM2G307119; Chuck et al., 2002) is involved in spikelet meristem identity in maize inflorescences, a phenotype that is affected in *ra3* mutants.

The *BES1* gene (GRMZM2G102514) regulates brassinosteroid-responsive gene expression and has been shown to interact with another class of TFs, *BES1 Interacting Myc-like* proteins (BIM), in Arabidopsis (Yin et al., 2005). Here, *BES1* and *BIM2* (GRMZM2G089501) were both significantly down-regulated in the *ra3* mutant. In addition, recent work has identified the Arabidopsis gene *SCHIZORIZA* (or *HSFB4* TF) as controlling patterning in stem cell divisions (ten Hove et al., 2010) and *AtMYB59* as playing a regulatory role in cell cycle progression (Mu et al., 2009). Putative orthologs of these TFs, GRMZM2G088242 and GRMZM2G130149, respectively, were misexpressed in the *ra3* mutant. In addition, differential expression of nine genes from the AP2/ERF family, including *BD1*, between the wild type and mutant suggests that these genes could be coregulated in response to a signal, such as ethylene, during this stage of development. Further analyses will test responses of these TFs and identify their targets in a panel of different mutants, developmental stages, and stress conditions.

Advantages of DGE for Genome-Wide Transcript Profiling

We found that DGE can be used to effectively determine genome-wide transcriptional changes. Aside from the prospect of using DGE in species where commercial arrays are not available, this method enables transcript profiling independent of prior knowledge of gene models, a considerable advantage over microarrays. This is especially relevant in cases where gene annotations are not complete. By mapping DGE tags and comparing abundances independently of gene models, we could resolve novel regions of expression in the maize genome. Some of these were specific to the *ra3* mutant, suggesting that analyses limited to existing gene models could exclude tissue- or mutant-specific transcripts. We also showed that even with short 20- to 21-nucleotide tags, we could delineate differential expression of closely related paralogs, which is limited in arrays due to cross-hybridization.

The DGE method also provides strand specificity, which is an advantage even over current RNA-seq protocols. Previous work showed that differential expression of antisense transcripts and S-AS pairs was common in maize (Ma et al., 2006). In human cell lines, evidence for shifts in the ratios of sense to antisense transcripts between normal and cancerous cells has indicated possible antisense-based regulation of developmental and disease processes (Chen et al., 2005; Morrissy et al., 2009). Consistent with results from other recent DGE studies, which resolved genome-wide expression of previously uncharacterized antisense transcripts (Morrissy et al., 2009; Babbitt et al., 2010), we identified a number of S-AS pairs with ratio changes between wild-type and *ra3* mutant samples. Further analyses will be needed to test the biological significance of the antisense transcripts; however, their detection provides a more accurate quantification of gene expression.

The ability to analyze strand-specific DGE data independent of gene models also enabled the detection and quantification of primary microRNA transcripts (pri-miRNAs). Based on computationally predicted miRNA hairpin structures (pre-miRNAs) and a set of PCR-RACE-validated pri-miRNA sequences identified by Zhang et al. (2009), we identified 14 pri-miRNAs (in 10 different families) represented by 10 or more reads in our DGE libraries (Supplemental Table S7). One abundant pri-miRNA, for miR159a, was significantly reduced in the *ra3* mutant; however, we observed no differential expression of its predicted target genes (Zhang et al., 2009), presumably due to miRNA redundancy. Resolution of miRNA abundances and correlations with expression differences in putative targets would be enhanced through the integration of parallel small RNA libraries with these data.

In addition, the tag-based nature of DGE technically generates only one read per transcript and thus improves cost efficiency by reducing redundancy. In this study, we multiplexed two enzyme-specific libraries in each lane of a flow cell using Illumina's first-phase GA. Although we achieved sufficient resolution of a large panel of TFs and differential expression profiles for low-frequency transcripts, the rapidly advancing technology currently generates estimated read depths of 3-fold higher than shown here. Given this current throughput, a multiplexed approach to concurrently sequence biological replicates and/or treatments would further reduce cost, time, and potential technical biases.

CONCLUSION

Digital gene expression profiling by high-throughput sequencing of 20- to 21-nucleotide tags revealed quantitative changes in transcript abundance on a genome-wide scale. Results from our DGE analyses in maize showed that we could effectively identify differentially expressed genes across a wide range of transcript abundances. Cumulative counts of tags that mapped to predicted gene models enabled the identification of functionally interesting genes and gene families with altered expression in *ra3* mutants. Our parallel analysis independent of gene models demonstrated that expression profiling was not limited by prior knowledge, thus promoting DGE as a platform for exploratory studies in species with nonsequenced genomes and for gene discovery. We also used the DGE data to resolve sense and antisense transcripts, distinguish between closely related paralogs, and identify unannotated genes and UTRs. Our approach used two enzymes to generate independent data sets that were multiplexed in a single lane of an Illumina flow cell. This provided a cost-effective method for orthogonal validation of genome-wide expression signatures and improved our coverage of the gene space. Our analyses, applications, and findings used here to interrogate the maize transcriptome and identify expression signatures underlying an agriculturally important trait are readily translated to other systems.

MATERIALS AND METHODS

DGE Library Construction and Sequencing

Field-grown maize (*Zea mays*) B73 and *ra3* plants were collected approximately 7 weeks after planting, and 2-mm ears were hand dissected and immediately frozen in liquid nitrogen. The *ra3* allele used here, *ra3-fea1*, results from an insertion, which leads to a frame shift. The *ra3* mutant was introgressed into B73 for five generations. RNA isolation, library construction, and sequencing were carried out at Pioneer Hi-Bred in Johnston, Iowa. Here, 500 to 2,000 ng of *DNaseI*-treated total RNA was used in library construction: double-stranded cDNAs were synthesized using oligo(dT) beads (Invitrogen). The cDNAs were then digested with an anchoring restriction enzyme (*Nla*III or *Dpn*II) and ligated to an Illumina-specific adapter, Adapter A, containing a recognition site for the type IIS restriction enzyme *Mme*I (New England Biolabs). Following *Mme*I digestion and dephosphorylation with shrimp alkaline phosphatase (USB Corp.), cDNAs were purified and a second Illumina adapter, Adapter B, containing a 2-bp degenerate 3′ overhang, was ligated. Tags flanked by both adapters were enriched by PCR using Phusion DNA polymerase (Finnzymes) and Gex PCR primers 1 and 2 (Illumina) following the manufacturer's instructions. The PCR products were run on a 12% PAGE gel, and the 85-bp DNA band was excised and purified using a Spin-X filter column (Costar) followed by ethanol precipitation. The DNA quality was assessed and quantified using an Agilent DNA 1000 series II assay, and the DNA sample was diluted to 10 nM. Cluster generation and sequencing were performed on the Illumina cluster station and Genome Analyzer (Illumina) following the manufacturer's instructions. Raw sequences were extracted from the resulting image files using the open source Firecrest and Bustard applications (Illumina). The *Nla*III reads were 21 bases long, since the enzyme cut site (CATG) overlapped with the *Mme*I binding site [TCC(G/A)AC] by 1 base in the sequencing primer. The additional base was later added to the 5′-end of each read in silico.

Mapping Pipeline and Extraction of Gene Information

We used Vmatch (www.vmatch.de) large-scale sequence analysis software to map the collapsed reads to the maize reference genome. The Vmktree feature was used to construct persistent indices for each of the 10 unmasked, assembled maize chromosomes and for an additional chromosome 0, which includes nonassembled sequence (B73 RefGen_v1). For each of the 11 indices, the unique *Dpn*II and *Nla*III tags were queried for complete matches to all possible 20- and 21-mer sequences, respectively, using the Vmatch algorithm. This included both sense and antisense matches, and strand information for each tag was retained. In phase II of the mapping pipeline, an editing distance of 1 was used to allow for a single mismatch, insertion, or deletion. Mapping data from all indices were parsed together, and results from *Dpn*II and *Nla*III libraries were analyzed independently. Tags that mapped uniquely up to three places in the maize genome were used to extract gene information using the Ensembl Application Perl Interface (http://uswest.ensembl.org/info/docs/api). A working gene set was used (gene build 4a.53; maizesequence.org), and for each gene model, the gene space was computationally extended by 300 nucleotides at both 5′- and 3′-ends to maximize the capture of complete UTRs. When we used a filtered set of 32,540 high-confidence maize gene models (build 4a.53; maizesequence.org), 22,700 genes were identified, of which 79% were found in both data sets and 6% and 14% were *Dpn*II and *Nla*III specific, respectively.

Strand Determination and Cumulative Read Counts

For each gene model, read counts associated with tags mapping in sense and antisense orientations were combined separately. Ratio changes for S-AS transcription of a given gene between the wild type and mutant were calculated based on analyses specified by Morrissy et al. (2009).

To determine consistency among libraries, genome-wide expression values (normalized to reads per million) were compared by pair-wise correlations of all libraries. Technical replicates of the same biological sample showed exceptional correlation in both *Dpn*II and *Nla*III libraries (Spearman r^2 = 0.998; Supplemental Fig. S5A). Combining all reads from multiple tags mapping to a given gene improved correlations among biological replicates

(r^2 = 0.963–0.861) compared with plotting individual tag frequencies alone (r^2 = 0.907–0.557; Supplemental Fig. S5B). Before determining a cumulative count per gene, we applied a scoring convention where a single read count was kept as 1 if a tag mapped to one place in the genome, 0.5 if it mapped to two locations, and 0.33 if it mapped three times.

Significance Testing

To compare gene expression profiles between wild-type and *ra3* mutant samples, we used the 26,663 and 30,746 genes identified from sense tags in phase I mapping of *Dpn*II and *Nla*III libraries, respectively. The edgeR package (www.bioconductor.org/packages/2.3/bioc/html/edgeR.html) adjusts for differences in library size; therefore, raw read counts per gene (or per tag) are directly used as input. One of the *Nla*III mutant samples, *ra3-3*, was highly variable when compared with all samples (Supplemental Fig. S5, B and C). This was most likely due to technical variation during library construction, since the *Dpn*II *ra3-3* correlated strongly with other biological replicates. After evaluation of the results from statistical tests, *Nla*III *ra3-3* was removed from the gene expression analyses presented here. We used a moderated, gene-wise dispersion analysis for both *Dpn*II and *Nla*III data sets separately with a weighted prior of 100. We applied a cutoff of nine reads for each gene, which reduced the set to 20,250 and 22,130 genes for *Dpn*II and *Nla*III libraries, respectively. Our significance threshold for differential expression was $P <$ 0.05 after correction using a Benjamini-Hochberg false discovery rate of 5%.

Quantitative Reverse Transcription-PCR Analyses

Total RNA samples for Q-PCR analyses represented three biological replicates of B73 and *ra3* ears and were comparable to those used for the DGE library construction. RNA integrity was assessed on an Agilent BioAnalyzer using a Nano Chip (Agilent 6000 Nano kit 5067-1511) according to the manufacturer's protocol. A total of four technical replicates were run for each RNA sample per assay using the AB7900 instrument (ABI; thermal cycling conditions were 50°C for 3 min [reverse transcription step], 95°C for 5 min [initial melt], and then 40 cycles of 94°C for 15 s and 60°C for 1 min). Specificity of each assay was determined by computer database homology searches. The linear dynamic range was determined using a standard curve generated from 1, 0.5, and 0.25 ng of RNA from a single replicate in each assay. The *EIF4a* reference gene was used as a normalization control and validated by correlation of its expression level (cycle threshold in 1-ng reactions) to the RNA concentration for all samples as determined from the Nanodrop quantifications. Sequences for forward primers (FP) and reverse primers (RP; Integrated DNA Technologies) and Taqman probes (5′-label, 6FAM; 3′-label, MGB [Applied Biosystems]) used to test genes by Q-PCR are as follows: *RA3* (GRMZM2G014729), FP: 5′-TGGACGAGCACAACAGCAA-3′, RP: 5′-AAGAAAACAACAAAAAAGGCCAGTA-3′, probe: 5′-AGGCGCTTATTAGCTACAA-3′; trehalose-phosphate synthase (GRMZM2G077659), FP: 5′-CTGGTGGTGAAAGGGTGGAT-3′, RP: 5′-GCTCTCCCAGATGCCGTAAG-3′, probe: 5′-CCCTGCTAGAGCCCCA-3′; TPP (GRMZM2G151044), FP: 5′-CGGCCGCACACAAAGC-3′, RP: 5′-GCGCCAACATGCTCAAAAC-3′, probe: 5′-CAGCGTCACTGAAAG-3′; Cyclin (GRMZM2G140633), FP: 5′-CGCCGGATTTCAACCAAA-3′, RP: 5′-TGGCTGTCTGCGCCTCTT-3′, probe: 5′-CGCCTGAAAGGCAA-3′; YABBY (similar to *Drooping Leaf*; GRMZM2G088309), FP: 5′-TGTACTTTTACCCCCGTACGTGTT-3′, RP: 5′-GGTGCGTACAATCCAACCATAA-3′, probe: 5′-CTGTTGCTGTTATTCTC-3′; NAM domain-containing (GRMZM2G054277), FP: 5′-ACTGGAGTACTCGATCCGCTTT-3′, RP: 5′-TGTAAGCTACGGCGGCAAA-3′, probe: 5′-CAACCTCGATCGCGATG-3′; *BES1* (GRMZM2G102514), FP: 5′-GCATTCGTGCTGAGTTTCGA-3′, RP: 5′-GCGTCACCTACGCCCTACA-3′, probe: 5′-CGGAGGCACATTC-3′; *IAA24* (GRMZM2G115357), FP: 5′-TCCATACATAAACAGAGGCTACAGACA-3′, RP: 5′-GATCCGTGTGTGCTCTTGGAT-3′, probe: 5′-CCACCTGGGAACGC-3′.

Detection of Pri-miRNAs

Locations of all DGE tags that mapped completely to the maize reference genome were compared with coordinates for a set of computationally predicated pre-miRNA hairpin structures (Zhang et al., 2009). Since the pre-miRNAs included only 250 nucleotides, an additional 500 nucleotides were computationally added to 5′- and 3′-ends of predicted hairpin sequences in order to simulate a region comparable to the length of a typical pri-miRNA and to recover tags within this space. Next, for each pre-miRNA identified, if it associated with a validated pri-miRNA model (Zhang et al., 2009), accurate mapping coordinates of the pri-miRNAs were incorporated to extend the search space as necessary and recover tags outside of the original 1,250-nucleotide window.

Sequence data from this article have been deposited in the National Center for Biotechnology Information Gene Expression Omnibus and are accessible through Gene Expression Omnibus Series accession number GSE24788.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Distribution of consensus tags and read counts.

Supplemental Figure S2. Alignment of one-mismatch reads to the genome.

Supplemental Figure S3. Mapping results comparing phase I with phase II mappings.

Supplemental Figure S4. Distribution of S-AS ratio changes.

Supplemental Figure S5. Pair-wise correlations between normalized DGE libraries.

Supplemental Figure S6. Top 10 most abundant tags sequenced in *Dpn*II and *Nla*III libraries.

Supplemental Figure S7. Smear plots from the edgeR-based analysis of gene expression.

Supplemental Figure S8. Regions of novel gene expression.

Supplemental Figure S9. Histogram of G+C content.

Supplemental Figure S10. Schematic of differentially expressed genes overlaid on pathways adapted from ricecyc.

Supplemental Table S1. Genes with S-AS ratio changes greater than 1.5-fold.

Supplemental Table S2. Genes with the most significant changes in transcript abundance.

Supplemental Table S3. All differentially expressed genes and their corresponding descriptions.

Supplemental Table S4. All putative TFs identified as differentially expressed in *Dpn*II and/or *Nla*III data sets.

Supplemental Table S5. Coordinates for individually mapped tags to AP2/ERF paralogs.

Supplemental Table S6. Genes associated with biochemical pathways based on ricecyc assignments.

Supplemental Table S7. miRNA precursors identified in the DGE data set.

Supplemental Data Set S1. All genes used in edgeR expression analysis and resulting fold changes and *P* values.

Supplemental Data Set S2. Individual signature tags used in edgeR expression analysis and resulting *P* values.

Supplemental Data Set S3. Maize genes identified as TFs based on sequence similarity to Arabidopsis genes.

ACKNOWLEDGMENTS

Special thanks to members of the Ware laboratory for helpful advice on computational analyses and on the manuscript, especially Jer-Ming Chia, Shiran Pasternak, and Andrew Olson. We also thank Peter van Buren for computer and database maintenance.

Received May 24, 2010; accepted August 23, 2010; published September 10, 2010.

LITERATURE CITED

Abouelhoda MI, Kurtz S, Ohlebusch E (2002) The enhanced suffix array and its applications to genome analysis. Algorithms in Bioinformatics, Proceedings **2452:** 449–463

Alexandrov NN, Brover VV, Freidin S, Troukhan ME, Tatarinova TV, Zhang HY, Swaller TJ, Lu YP, Bouck J, Flavell RB, et al (2009) Insights into corn genes derived from large-scale cDNA sequencing. Plant Mol Biol **69:** 179–194

Amit I, Garber M, Chevrier N, Leite AP, Donner Y, Eisenhaure T, Guttman M, Grenier JK, Li WB, Zuk O, et al (2009) Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. Science **326:** 257–263

Asmann YW, Klee EW, Thompson EA, Perez EA, Middha S, Oberg AL, Therneau TM, Smith DI, Poland GA, Wieben ED, et al (2009) 3′ tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. BMC Genomics **10:** 531

Babbitt CC, Fedrigo O, Pfefferle AD, Boyle AP, Horvath JE, Furey TS, Wray GA (2010) Both noncoding and protein-coding RNAs contribute to gene expression evolution in the primate brain. Genome Biol Evol **2:** 67–79

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. Plant J **51:** 910–918

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, et al (2009) NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res **37:** D885–D890

Bevan M, Walsh S (2005) The Arabidopsis genome: a foundation for plant research. Genome Res **15:** 1632–1642

Blencowe BJ, Ahmad S, Lee LJ (2009) Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. Genes Dev **23:** 1379–1386

Blow N (2009) Transcriptomics: the digital generation. Nature **458:** 239–242

Brady SM, Provart NJ (2009) Web-queryable large-scale data sets for hypothesis generation in plant biology. Plant Cell **21:** 1034–1051

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo SJ, McCurdy S, Foy M, Ewan M, et al (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol **18:** 630–634

Capaldi AP, Kaplan T, Liu Y, Habib N, Regev A, Friedman N, O'Shea EK (2008) Structure and function of a transcriptional network activated by the MAPK Hog1. Nat Genet **40:** 1300–1306

Chen JJ, Sun M, Hurst LD, Carmichael GG, Rowley JD (2005) Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. Trends Genet **21:** 326–329

Cho YH, Yoo SD, Sheen J (2006) Regulatory functions of nuclear hexokinase1 complex in glucose signaling. Cell **127:** 579–589

Chuck G, Muszynski M, Kellogg E, Hake S, Schmidt RJ (2002) The control of spikelet meristem identity by the branched silkless1 gene in maize. Science **298:** 1238–1241

Emrich SJ, Barbazuk WB, Li L, Schnable PS (2007a) Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res **17:** 69–73

Emrich SJ, Li L, Wen TJ, Yandeau-Nelson MD, Fu Y, Guo L, Chou HH, Aluru S, Ashlock DA, Schnable PS (2007b) Nearly identical paralogs: implications for maize (Zea mays L.) genome evolution. Genetics **175:** 429–439

Eveland AL, McCarty DR, Koch KE (2008) Transcript profiling by 3′-untranslated region sequencing resolves expression of gene families. Plant Physiol **146:** 32–44

Fahlgren N, Sullivan CM, Kasschau KD, Chapman EJ, Cumbie JS, Montgomery TA, Gilbert SD, Dasenko M, Backman TWH, Givan SA, et al (2009) Computational and analytical framework for small RNA profiling by high-throughput sequencing. RNA **15:** 992–1002

Flicek P, Aken BL, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Coates G, Fairley S, et al (2010) Ensembl's 10th year. Nucleic Acids Res **38:** D557–D562

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge YC, Gentry J, et al (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol **5:** R80

Goda H, Sasaki E, Akiyama K, Maruyama-Nakashita A, Nakabayashi K, Li WQ, Ogawa M, Yamauchi Y, Preston J, Aoki K, et al (2008) The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. Plant J **55:** 526–542

Gregory BD, Yazaki J, Ecker JR (2008) Utilizing tiling microarrays for whole-genome analysis in plants. Plant J **53:** 636–644

Guo M, Yang S, Rupe M, Hu B, Bickel DR, Arthur L, Smith O (2008) Genome-wide allele-specific expression analysis using massively parallel signature sequencing (MPSS) reveals cis- and trans-effects on gene expression in maize hybrid meristem tissue. Plant Mol Biol **66:** 551–563

Gutiérrez RA, Lejay LV, Dean A, Chiaromonte F, Shasha DE, Coruzzi GM (2007) Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis. Genome Biol **8:** R7

Gutiérrez RA, Stokes TL, Thum K, Xu X, Obertello M, Katari MS, Tanurdzic M, Dean A, Nero DC, McClung CR, et al (2008) Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene CCA1. Proc Natl Acad Sci USA **105:** 4939–4944

Harbers M, Carninci P (2005) Tag-based approaches for transcriptome research and genome annotation. Nat Methods **2:** 495–502

Horan K, Jang C, Bailey-Serres J, Mittler R, Shelton C, Harper JF, Zhu JK, Cushman JC, Gollery M, Girke T (2008) Annotating genes of known and unknown function by large-scale coexpression analysis. Plant Physiol **147:** 41–57

Jongeneel CV, Iseli C, Stevenson BJ, Riggins GJ, Lal A, Mackay A, Harris RA, O'Hare MJ, Neville AM, Simpson AJG, et al (2003) Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. Proc Natl Acad Sci USA **100:** 4702–4705

Kaufmann K, Muiño JM, Jauregui R, Airoldi CA, Smaczniak C, Krajewski P, Angenent GC (2009) Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the Arabidopsis flower. PLoS Biol **7:** e1000090

Kolbe A, Tiessen A, Schluepmann H, Paul M, Ulrich S, Geigenberger P (2005) Trehalose 6-phosphate regulates starch synthesis via posttranslational redox activation of ADP-glucose pyrophosphorylase. Proc Natl Acad Sci USA **102:** 11118–11123

Levesque MP, Vernoux T, Busch W, Cui HC, Wang JY, Blilou I, Hassan H, Nakajima K, Matsumoto N, Lohmann JU, et al (2006) Whole-genome analysis of the SHORT-ROOT developmental pathway in Arabidopsis. PLoS Biol **4:** e143

Li H, Lovci MT, Kwon YS, Rosenfeld MG, Fu XD, Yeo GW (2008) Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. Proc Natl Acad Sci USA **105:** 20179–20184

Liang CZ, Jaiswal P, Hebbard C, Avraham S, Buckler ES, Casstevens T, Hurwitz B, McCouch S, Ni JJ, Pujar A, et al (2008) Gramene: a growing plant comparative genomics resource. Nucleic Acids Res **36:** D947–D953

Lister R, Gregory BD, Ecker JR (2009) Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. Curr Opin Plant Biol **12:** 107–118

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell **133:** 523–536

Ma J, Morrow DJ, Fernandes J, Walbot V (2006) Comparative profiling of the sense and antisense transcriptome of maize lines. Genome Biol **7:** R22

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res **18:** 1509–1517

Messing J, Dooner HK (2006) Organization and variability of the maize genome. Curr Opin Plant Biol **9:** 157–163

Metzker ML (2010) Sequencing technologies: the next generation. Nat Rev Genet **11:** 31–46

Meyers BC, Tej SS, Vu TH, Haudenschild CD, Agrawal V, Edberg SB, Ghazal H, Decola S (2004) The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. Genome Res **14:** 1641–1653

Morrissy AS, Morin RD, Delaney A, Zeng T, McDonald H, Jones S, Zhao Y, Hirst M, Marra MA (2009) Next-generation tag sequencing for cancer gene expression profiling. Genome Res **19:** 1825–1835

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods **5:** 621–628

Mu RL, Cao YR, Liu YF, Lei G, Zou HF, Liao Y, Wang HW, Zhang WK, Ma B, Du JZ, et al (2009) An R2R3-type transcription factor gene AtMYB59 regulates root growth and cell cycle progression in Arabidopsis. Cell Res **19:** 1291–1304

Nobuta K, Venu RC, Lu C, Beló A, Vemaraju K, Kulkarni K, Wang WZ, Pillay M, Green PJ, Wang GL, et al (2007) An expression atlas of rice mRNAs and small RNAs. Nat Biotechnol **25:** 473–477

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet **40:** 1413–1415

Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, et al (2009) ArrayExpress update: from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res **37:** D868–D872

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature **464:** 768–772

Pruneda-Paz JL, Breton G, Para A, Kay SA (2009) A functional genomics approach reveals CHE as a component of the Arabidopsis circadian clock. Science **323:** 1481–1485

Ramsey SA, Klemm SL, Zak DE, Kennedy KA, Thorsson V, Li B, Gilchrist M, Gold ES, Johnson CD, Litvak V, et al (2008) Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics. PLoS Comput Biol **4:** e1000021

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26:** 139–140

Robinson MD, Smyth GK (2007) Moderated statistical tests for assessing differences in tag abundance. Bioinformatics **23:** 2881–2887

Robinson MD, Smyth GK (2008) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics **9:** 321–332

Rolland F, Baena-Gonzalez E, Sheen J (2006) Sugar sensing and signaling in plants: conserved and novel mechanisms. Annu Rev Plant Biol **57:** 675–709

Satoh-Nagasawa N, Nagasawa N, Malcomber S, Sakai H, Jackson D (2006) A trehalose metabolic enzyme controls inflorescence architecture in maize. Nature **441:** 227–230

Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science **326:** 1112–1115

Siddiqui AS, Delaney AD, Schnerch A, Griffith OL, Jones SJM, Marra MA (2006) Sequence biases in large scale gene expression profiling data. Nucleic Acids Res **34:** e83

Simon SA, Zhai JX, Nandety RS, McCormick KP, Zeng J, Mejia D, Meyers BC (2009) Short-read sequencing technologies for transcriptional analyses. Annu Rev Plant Biol **60:** 305–333

Smith AM, Zeeman SC, Smith SM (2005) Starch degradation. Annu Rev Plant Biol **56:** 73–98

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, et al (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science **321:** 956–960

Tang FC, Barbacioru C, Wang YZ, Nordman E, Lee C, Xu NL, Wang XH, Bodeau J, Tuch BB, Siddiqui A, et al (2009) mRNA-seq whole-transcriptome analysis of a single cell. Nat Methods **6:** 377–382

ten Hove CA, Willemsen V, de Vries WJ, van Dijken A, Scheres B, Heidstra R (2010) SCHIZORIZA encodes a nuclear factor regulating asymmetry of stem cell divisions in the Arabidopsis root. Curr Biol **20:** 452–457

’t Hoen PAC, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RH, de Menezes RX, Boer JM, van Ommen GJB, den Dunnen JT (2008) Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res **36:** e141

Toufighi K, Brady SM, Austin R, Ly E, Provart NJ (2005) The Botany Array Resource: e-northerns, expression angling, and promoter analyses. Plant J **43:** 153–163

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science **270:** 484–487

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res **19:** 327–335

Vollbrecht E, Springer PS, Goh L, Buckler ES IV, Martienssen R (2005) Architecture of floral branch systems in maize and related grasses. Nature **436:** 1119–1126

Wang ET, Sandberg R, Luo SJ, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. Nature **456:** 470–476

Wang XF, Elling AA, Li XY, Li N, Peng ZY, He GM, Sun H, Qi YJ, Liu XS, Deng XW (2009) Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in maize. Plant Cell **21:** 1053–1069

Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB (2007) Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. Plant Physiol **144:** 32–42

Wei F, Coe E, Nelson W, Bharti AK, Engler F, Butler E, Kim H, Goicoechea JL, Chen M, Lee S, et al. (2007) Physical and genetic structure of the maize genome reflects its complex evolutionary history. PLoS Genet **3:** e123

Yin YH, Vafeados D, Tao Y, Yoshida S, Asami T, Chory J (2005) A new class of transcription factors mediates brassinosteroid-regulated gene expression in Arabidopsis. Cell **120:** 249–259

Zhang LF, Chia JM, Kumari S, Stein JC, Liu ZJ, Narechania A, Maher CA, Guill K, McMullen MD, Ware D (2009) A genome-wide characterization of microRNA genes in maize. PLoS Genet **5:** e1000716

Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W (2004) GENEVESTIGATOR: Arabidopsis microarray database and analysis toolbox. Plant Physiol **136:** 2621–2632